**Supplementary Materials**

Table of Contents

[Supplementary Table 1: Definitions of the metrics used in the simulation 2](#_Toc73765863)

[Supplementary Table 2: Variables in the Tehran housing data 3](#_Toc73765864)

[Caption: Supplementary Figure 1 4](#_Toc73765865)

[Caption: Supplementary Figure 2 4](#_Toc73765866)

[Caption: Supplementary Figure 3 4](#_Toc73765867)

[Caption: Supplementary Figure 4 4](#_Toc73765868)

[Caption: Supplementary Figure 5 4](#_Toc73765869)

[Caption: Supplementary Figure 6 4](#_Toc73765870)

[Caption: Supplementary Figure 7 5](#_Toc73765871)

[Caption: Supplementary Figure 8 5](#_Toc73765872)

[Caption: Supplementary Figure 9 5](#_Toc73765873)

Let denote the total number of variables, denote the number of true signals and denote the number of true noise variables. Suppose an algorithm identifies features as signals, of which only are correctly identified as signals. The remaining signals, if any, are mistakes. Some of the features that were called noise may have been signals too. Below are the formal definitions.

# Supplementary Table 1: Definitions of the metrics used in the simulation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Truth** | |  |
|  |  | **Signals** | **Noise** | **Subtotal** |
| **Algorithm** | **Signals** | v | r-v | r |
| **Noise** | s-v | p-r-s+v | p-r |
|  | **Subtotal** | s | p-s | p |

* Type I error rate (proportion of noise variables that were claimed to be signals)

or P(Algorithm claims feature is a signal | feature is noise) = (r-v)/(p-s)

* Power (proportion of true signals that were correctly selected)

or P(Algorithm claims feature is a signal | feature is signal) = v/s

* pFDR (proportion of algorithm identified signals that are really noise features)

or P(feature is noise | Algorithm claims feature is a signal) = (r-v)/r

* pFNR (proportion of algorithm identified noise variables that are really true signals)

or P(feature is signal | Algorithm claims feature is a noise) = (s-v)/(p-r)

# Supplementary Table 2: Variables in the Tehran housing data

|  |  |  |
| --- | --- | --- |
| **Type** | **Variable name** | **Variable label** |
| Outcome | V9 | Actual sales price |
| Independent variables:  project physical and financial features | V2 | Total floor area of the building |
| V3 | Lot area |
| V4 | Total Preliminary estimated construction cost based on the prices at the beginning of the project |
| V5 | Preliminary estimated construction cost based on the prices at the beginning of the project |
| V6 | Equivalent preliminary estimated construction cost based on the prices at the beginning of the project in a selected base year |
| V7 | Duration of construction |
| V8 | Price of the unit at the beginning of the project per square meter |
| Independent variables:  economic variables and indices | V11 | The number of building permits issued |
| V12 | Building services index for preselected base year |
| V13 | Wholesale price index of building materials for the base year |
| V14 | Total floor areas of building permits issued by the city/municipality |
| V15 | Cumulative liquidity |
| V16 | Private sector investment in new buildings |
| V17 | Land price index for the base year |
| V18 | The number of loans extended by banks in a time resolution |
| V19 | The amount of loans extended by banks in a time resolution |
| V20 | The interest rate for loan in a time resolution |
| V21 | The average construction cost by private sector at the completion of construction |
| V22 | The average cost of buildings by private sector at the beginning of construction |
| V23 | Official exchange rate with respect to dollars |
| V24 | Nonofficial (street market) exchange rate with respect to dollars |
| V25 | Consumer price index (CPI) in the base year |
| V26 | CPI of housing, water, fuel & power in the base year |
| V27 | Stock market index |
| V28 | Population of the city |
| V29 | Gold price per ounce |

The data were acquired from University of California, Irvine Machine Learning Repository at <http://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set> on October 24, 2020.

# Caption: Supplementary Figure 1

Sensitivity of the null bound in the ProSGPV algorithm. The size of the null bound and the capture rate of the exact true model from ProSGPV with different null bounds are compared under medium and high signal-to-noise ratios. Choices of the null bound include the original bound , , ,  , and 0. The top plot shows how the null bound (median, first and third quartiles) changes with , and the bottom plot shows the capture rates surrounded by 95% Wald confidence intervals over 1000 simulations.

# Caption: Supplementary Figure 2

Support recovery rates of one-stage and two-stage ProSGPV algorithms. Each curve represents the capture rate of the exact true model over 1000 simulations. The two-stage algorithm selects the that minimizes the generalized information criterion in the first stage. Shaded belts are 95% Wald confidence intervals.

# Caption: Supplementary Figure 3

Estimated power and Type I Error rates of all algorithms under combinations of autocorrelation level, signal-to-noise ratios, and (). Rates are plotted against either the ratio of the sample size over the number of variables , or only.

# Caption: Supplementary Figure 4

False discovery rate (FDR) and false non-discovery rate (FNDR) of all algorithms under combinations of autocorrelation level, signal-to-noise ratios and (n, p, s). Rates are plotted against either the ratio of sample size over the number of variables, or the number of variables.

# Caption: Supplementary Figure 5

How different parameter tuning methods affects MC+. Support recovery rates, parameter estimation error measured by mean absolute error, and prediction accuracy measured by prediction root mean square error in an independent test set are compared for MC+ implemented in two ways. One way is to select the that minimizes the generalized information criterion. The other way is to use a universal .

# Caption: Supplementary Figure 6

Comparison of computation costs of all algorithms under combinations of autocorrelation level, signal-to-noise ratios, and (n, p, s). Solid lines are the median running time and the shades are first and third quartiles of the running time. For aesthetic reasons, values are censored at 0.5 second.

# Caption: Supplementary Figure 7

Clustering and correlation patterns of the Tehran housing data. The location of variables represents their clustering pattern. Each pair of variables is connected with a curve whose color represents the correlation level: strong positive correlation is in purple while strong negative correlation is in red.

# Caption: Supplementary Figure 8

Histograms of model size from each algorithm in the training set over 1000 repetitions. The high SNR case includes all 26 variables in the Tehran housing data, where the in the full OLS model; the medium SNR case includes only nine variables, where the in the reduced OLS model.

# Caption: Supplementary Figure 9

Boxplots of prediction root mean square errors from each algorithm in the test set over 1000 repetitions. The high SNR case includes all 26 variables in the Tehran housing data, where the in the full OLS model; the medium SNR case includes only nine variables, where the in the reduced OLS model.